YOUKU-MPLUG: A 10 MILLION LARGE-SCALE CHINESE VIDEO-LANGUAGE PRE-TRAINING DATASET AND BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We firstly release the largest public Chinese high-quality video-language dataset named Youku-mPLUG, which is collected from Youku[1], a well-known Chinese video-sharing website, with strict criteria of safety, diversity, quality, and copyright. Youku-mPLUG contains 10 million Chinese video-text pairs filtered from 400 million raw videos across a wide range of 45 diverse categories for large-scale pre-training. In addition, to facilitate a comprehensive evaluation of video-language models, we carefully build the largest human-annotated Chinese benchmarks covering three popular video-language tasks across cross-modal retrieval, video captioning, and video category classification. We also provide comprehensive benchmark evaluations of models across different architectures including encoder-only (i.e., ALPRO), encoder-decoder (i.e., mPLUG-2), and decoder-only (i.e., mPLUG-Video) for comparison. Especially, we train the first Chinese Multimodal LLM with only 1.7% trainable parameters for video understanding. Experiments show that models pre-trained on Youku-mPLUG gain up to 23.1% improvement in video category classification. Besides, mPLUG-video achieves a new state-of-the-art result on these benchmarks with 80.5% top-1 accuracy in video category classification and 68.9 CIDEr score in video captioning, respectively. Finally, the 2.7B version of mPLUG-video demonstrates impressive instruction and video understanding ability. The zero-shot instruction understanding experiment indicates that pretraining with Youku-mPLUG can enhance the ability to comprehend overall and detailed visual semantics, recognize scene text, and leverage open-domain knowledge.

1 INTRODUCTION

With the release of large-scale English video-language datasets (e.g., Howto100M(Miech et al., 2019) and WebVid-2.5M(Bain et al., 2021)), video-language pre-training (VLP) has achieved the superior performance on various downstream tasks, such as video-text retrieval, video question answering, and video captioning. Moreover, the recent multimodal LLM in video (e.g., VideoChat(Li et al., 2023b), Flamingo(Alayrac et al., 2022)) has demonstrated strong zero-shot video understanding ability based on these large-scale datasets. Compared with the English VLP community as Tab. 1, the lack of large-scale and high-quality public Chinese VLP datasets hinders the research of Chinese video-language pretraining and multimodal LLM. In addition, publicly available benchmarks as Tab. 2 are also missing for the Chinese VLP community. These limitations will result in two significant issues. Firstly, the development and application of Chinese VLP and multimodal LLM are being lagged behind. Secondly, the comparison between different methods becomes challenging due to the fairness issue that some works are able to achieve surprisingly good performance by using secret downstream benchmarks. While some methods translate English text into Chinese (Madasu et al., 2022) or annotate the dataset based on the English video (Wang et al., 2019), there remains an intrinsic linguistic and cultural gap between English and Chinese.

To facilitate the research and application of Chinese VLP, we release the first and largest public Chinese video-language pretraining dataset and benchmarks named Youku-mPLUG, which is collected from Youku, a well-known Chinese video-sharing website with strict criteria of safety, diversity,

[1]https://www.youku.com

Table 1: Statistics of Youku-mPLUG and its comparison with existing video-language pre-training datasets.

Dataset Name	Language	# Videos	# Text	Avg. Len (secs)	Duration (hrs)	Domain	Availability
HowTo100M (Miech et al., 2019)	English	136M	136M	3.6	135K	Instruction	✓
YT-Temporal-180M (Zellers et al., 2021)	English	180M	180M	-	-	Instruction	✓
HD-VILA-100M (Xue et al., 2022)	English	103M	103M	13.4	372K	Open	✓
WebVid10M (Bain et al., 2021)	English	10M	10M	18.0	52K	Open	✓
ALIVOL-10M (Lei et al., 2021a)	Chinese	103M	110M	34.6	99K	E-Commerce	✗
Kwai-SVC-11M (Nie et al., 2022b)	Chinese	11M	4M	57.9	177K	Open	✗
CREATE-10M (Zhang et al., 2022)	Chinese	10M	10M	29.8	83K	Open	✗
CNVid-3.5M (Gan et al., 2023)	Chinese	3.5M	3.5M	36.2	35K	Open	✗
Youku-mPLUG	Chinese	10M	10M	54.2	150K	Open	✓

Table 2: Statistics of Youku-mPLUG and its comparison with existing video-language downstream datasets.

Dataset Name	Language	# Sample	Domain	Retrieval	Classification	Caption	Availability
MSRVTT (Xu et al., 2016)	English	10K	Open	✓	✓	✓	✓
DiDeMo (Anne Hendricks et al., 2017)	English	27K	Flickr	✓	✗	✗	✗
MSVD (Chen & Dolan, 2011)	English	10K	Open	✓	✓	✓	✓
LSMDC (Rohrbach et al., 2015)	English	118K	Movie	✓	✓	✗	✓
ActivityNet (Krishna et al., 2017)	English	100K	Open	✓	✓	✗	✓
VATEX (Wang et al., 2019)	English/Chinese	41K	Kinetics-600	✓	✗	✓	✓
BFVD (Zhang et al., 2020)	Chinese	43K	E-Commerce	✓	✗	✗	✗
FFVD (Zhang et al., 2020)	Chinese	32K	E-Commerce	✓	✗	✗	✗
CREATE-210K (Zhang et al., 2022)	Chinese	216K	Open	✓	✗	✓	✗
Youku-mPLUG	Chinese	365K	Open	✓	✓	✓	✓

quality and copyright. Youku-mPLUG contains 10 million video-text pairs for pre-training and 0.3 million videos for downstream benchmarks. For the pre-training dataset, we collect 10 million high-quality video-text pairs filtered from 400 million raw videos with the strict criteria of safety, diversity, and quality. **Safety**, the dataset is subject to heavy filtering and restrictions through an in-house multi-level risk detection system to prevent any content related to high risks; **Diversity**, the videos are carefully classified into 45 diverse categories covering various domains, e.g., Daily life, Comedy, and Pet, with a balanced distribution; **Quality**, we have conducted strict data cleaning at both the text and video levels, while using Chinese image-text pre-trained model to improve the data quality. Furthermore, We build the largest human-annotated Chinese benchmarks covering Cross-modal Retrieval, Video Captioning, and Video Category Classification for comprehensive evaluation of video-language models and downstream applications. For each downstream task, we hire well-educated people and adopt a two-step verification to ensure the quality and diversity of the annotations In concrete, We would first hire a group of well-educated people to annotate a small fraction of data with provided annotation details and instructions. Then we scrutinize the annotated data and filter out those annotators who have extremely poor annotation quality. We also revised the annotation instructions according to the problems during the first-round annotation. After that, we give another small fraction of the data for annotation. If the quality of these annotations meets the requirement, we would provide all of the data for labeling. Otherwise, we repeat the previous checking procedure.

Besides, we investigate popular video-language models, the encoder-only model ALPRO (Li et al., 2022b) and the encoder-decoder model mPLUG-2 (Xu et al., 2023) pre-trained on Youku-mPLUG. Drawing inspiration from the idea of modularization (Li et al., 2022a; Xu et al., 2023; Ye et al., 2023), we propose the modularized decoder-only model mPLUG-video with limited trainable parameters, which consists of the trainable video encoder, visual abstractor module, and the frozen pre-trained LLM decoder. We first obtain dense video representations from the video encoder. Then, we employ the visual abstractor module to summarize visual information with several learnable tokens. Finally, the visual representations are combined with text queries and fed into the frozen LLM decoder to generate the response. Experiments show that models pre-trained on Youku-mPLUG gain up to 23.1% improvement in video category classification. With the proposed dataset, mPLUG-video achieves 80.5% top-1 accuracy in video category classification and 68.9 CIDEr score in video captioning, respectively. It becomes new state-of-the-art results on these benchmarks. Moreover, we scale up mPLUG-video based on frozen Bloomz(Workshop et al., 2023) as Chinese multimodal LLM with only 1.7% trainable parameters, which demonstrates impressive instruction and video

understanding ability. As an insight, our zero-short video instruction understanding test validates that Youku-mPLUG can strengthen the scene text recognizing ability and incorporate open-domain knowledge for video understanding. Qualitative results can be found in the Supplementary Material. These pre-trained models have also been released to facilitate the research and application of Chinese video-language pre-training.

In summary, our main contributions are:

- We release the first and largest Chinese video-language pretraining dataset and benchmarks named Youku-mPLUG.
- We provide comprehensive benchmark evaluations of models across different architectures including encoder-only (i.e., ALPRO), encoder-decoder (i.e., mPLUG-2), and our proposed modularized decoder-only mPLUG-video pre-trained on Youku-mPLUG for comparison.
- We scale up and release mPLUG-video based on Bloomz as Chinese multimodal LLM with only 1.7% trainable parameters, which demonstrates the impressive zero-shot instruction and video understanding ability.
- Experiments show that models pre-trained on Youku-mPLUG gain a significant improvement over baselines and mPLUG-video achieves state-of-the-art results on these benchmarks.

2 Related Work

Video-Language Pre-training Datasets Large-scale datasets have proven effective for video-language representation learning. Previously, most video-language models were trained on the HowTo100M dataset (Miech et al., 2019), which comprises 136 million video clips from 1.22 million instructional YouTube videos. However, this dataset is limited to the instructional domain and is unsuitable for generalization. To overcome this constraint, Zeller et al. (Zellers et al., 2021) and Xue et al. (Xue et al., 2022) propose the YT-Temporal-180M and HD-VILA-100M corpus, respectively. Meanwhile, to reduce the noise in subtitles, Bain et al. (Bain et al., 2021) introduce the Webvid10M dataset which is inspired by the collection schemes of Conceptual Caption datasets (Sharma et al., 2018). However, these datasets are limited to English language corpus and cannot be directly applied to the Chinese domain. Although there exist some large-scale Chinese video-language datasets such as ALIVOL (Lei et al., 2021a), Kwai-SVC (Nie et al., 2022a), CREATE-10M (Zhang et al., 2022), and CNVid-3.5M (Gan et al., 2023), none of them have been publicly released to date, which hinders the progress of research in the Chinese video-language learning field. To address this gap, we present Youku-mPLUG, the largest Chinese high-quality video-language dataset, to facilitate future research on large-scale video-language learning in the Chinese language.

Video-Language Downstream Benchmarks For evaluating video-language pre-training models, researchers have proposed several downstream tasks such as video-text retrieval, video question answering, and video captioning for performance evaluation. For instance, MSRVTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), and LSMDC (Rohrbach et al., 2015) are commonly adopted for text-video retrieval evaluation. Similarly, MSRVTT-QA (Xu et al., 2017), MSVD-QA (Xu et al., 2017), and T-GIF (Jang et al., 2017) are widely used for video question evaluation. Meanwhile, MSRVTT-Caption (Xu et al., 2016) and MSVD-Caption (Chen & Dolan, 2011) are commonly used for video caption evaluation. However, these datasets are primarily collected from YouTube, which is not entirely suitable for the Chinese domain. Furthermore, while there are some Chinese benchmark datasets such as CREATE (Zhang et al., 2022) and VATEX (Wang et al., 2019), they are not fully released and only evaluate one aspect of the model's performance. Additionally, there is a lack of systematic video language downstream benchmarks or leaderboards for Chinese video-language pre-training evaluation. Consequently, we propose three downstream benchmarks, including video category classification, video-text retrieval, and video captioning, for evaluating models' performance on Youku-mPLUG. These benchmarks are specifically designed for the Chinese domain and are intended to fill the gap in existing English benchmarks, which may not be entirely suitable for Chinese video-language pre-training evaluation.

Video-Language Pre-training Models In recent years, there has been a growing interest in video-language pre-training, and various methods have been proposed to explore this area. Traditional approaches (Luo et al., 2020; Li et al., 2020) rely on pre-extracted, dense video frame or clip features for video-language representation. In contrast, ClipBERT (Lei et al., 2021b) introduces a sparse sampling strategy that facilitates end-to-end learning while simultaneously improving performance.

Building upon this strategy, many approaches (Bain et al., 2021; Ge et al., 2022) have been developed, which incorporate novel architectures and pre-training tasks for video-language learning. For example, Frozen (Bain et al., 2021) and BridgeFormer (Ge et al., 2022) employ contrastive learning to align the semantics of paired video and text in the same embedding space. Additionally, ALPRO (Li et al., 2022b), TW-BERT (Yang et al., 2023), mPLUG-2 (Xu et al., 2023), and HiTeA (Ye et al., 2022) fuse video and language features to generate video-language representations for understanding and generation. Recently, large language models such as GPT-3 (Brown et al., 2020), Bloom (Workshop et al., 2023), and LLaMA (Touvron et al., 2023) have demonstrated significant zero-shot generalization abilities, which are advantageous for the vision-language field. For instance, BLIP-2 (Li et al., 2023a), miniGPT-4 (Zhu et al., 2023), and mPLUG-Owl (Ye et al., 2023) exhibit robust zero-shot generalization and conversation capabilities by aligning vision and language models. In this work, we provide a decoder-only video-language model mPLUG-video pre-trained on our Youku-mPLUG dataset with a strong generalization performance in terms of both video-language understanding and generation.

3 YOUKU-MPLUG DATASET CREATION

To fill in the blank of the public Chinese video-text pre-training dataset and benchmarks, we release the largest public Chinese Video-language dataset named Youku-mPLUG collected with the strict criteria of safety, diversity, and quality from Youku, a Chinese video-sharing website. Youku-mPLUG contains 10 million video-text pairs for pre-training and 0.3 million videos for downstream benchmarks covering Video-Text Retrieval, Video Captioning, and Video Category Classification. Randomly sampled examples are shown in Figure 1.

Figure 1: Random sampled examples in Youku-mPLUG.

3.1 PRE-TRAINING DATASET CONSTRUCTION

For the pre-training dataset, we filter 10 million high-quality video-text pairs from 400 million raw videos with strict safety, diversity, and quality criteria. In terms of safety, the dataset is heavily filtered and restricted by an internal multi-level risk detection system with both multimodal model detection and manual review processes to prevent any content related to pornography, violence, terrorism, discrimination, abuse, or other high risks. In specific, the safety detection system primarily consists of two components. Firstly, we utilize in-house visual and language models to identify potentially hazardous content in videos and title information, including pornography, violence, terrorism, discrimination, abuse, etc., and ensemble the results. Secondly, a crowd-sourcing platform is employed for manual re-checking, in cases where it is challenging for the models to differentiate (e.g., when scores are indistinguishable). The annotation results will be fed to the model for more refined training. Regarding diversity, we have applied video fingerprinting technology to eliminate videos that are completely identical. With the hierarchical multi-label classification model (Giunchiglia & Lukasiewicz, 2020), the videos are carefully classified into 20 super categories and

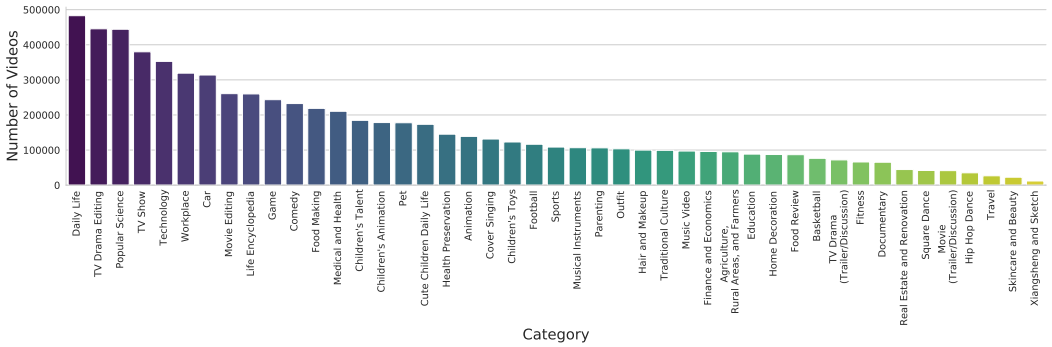


Figure 2: The distribution of the number of videos in each common category.

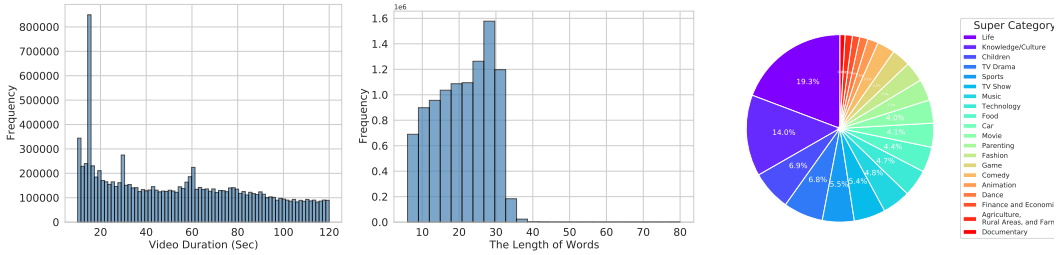


Figure 3: Youku-mPLUG dataset statistics: we report the histogram of video duration in seconds (left), the histogram of title length in words (middle), and the ratios of the categories in each super-category (right).

45 common categories as Fig. 2, covering various domains, with a balanced distribution. To ensure high quality, we have conducted strict data cleaning at both the text and video levels. For text, we have imposed language restrictions on video titles, requiring the length to be between 5 and 30 words and including at least 5 Chinese characters while filtering out those with obvious advertising or meaningless content. In terms of video quality and integrity, we have specifically chosen recently uploaded videos with durations ranging from 10 to 120 seconds to ensure clear and complete content. Further, we also employ the Chinese image-text pre-trained model CLIP (Yang et al., 2022) to improve the data quality by deprecating those with low similarities between the mean frame features and text features. Fig. 3 shows the statistics of video duration and word length. Furthermore, to safeguard the copyright of videos, we manually insert a 2-second shallow watermark at the start of each video, which is indispensable to open-source these videos. As demonstrated in (Bain et al., 2021), these watermarks do not impact the performance of the model.

3.2 Downstream Benchmark Construction

For the downstream benchmark, we design three types of tasks including video-text retrieval, video category classification, and video captioning to evaluate the performance in terms of understanding and generation. The statistics of these three different datasets are summarized in Tab. 3.

Table 3: Statistics of Youku-mPLUG benchmark datasets. # pairs indicates the number of video-text pairs.

Task	Train (# Pairs)	Val (# Pairs)	Test (# Pairs)
Video Category Classification	100,023	14,678	20,026
Video-Text Retrieval	37,595	7,271	7,414
Video Captioning	170,866	7,510	7,705

Video Category Classification Our initial step involves randomly selecting a substantial number of videos based on category frequency. Next, we collect the video categories from the Youku database, which are auto-generated by an online model. It is important to note that this model's accuracy is approximately 94% when considering historical prediction data, thus not entirely reliable.

Consequently, we put forth additional efforts to ensure the quality and accuracy of our datasets by manually verifying each video and its corresponding title in the benchmark datasets. Prior to annotation, we supply a smaller dataset containing 100 videos, along with their metadata, including titles and categories generated by the online prediction model. Annotators are then tasked with confirming the assigned categories in relation to the videos and their titles. They must also assign a relevance score, which ranges from 1 to 5. A higher score suggests a greater likelihood of the video belonging to the given category, and those with scores above 3 are retained. Annotators with error rates exceeding 2.5% are disqualified. After eliminating unsuitable annotators, we proceed with annotating the video category classification dataset. To ensure the utmost accuracy, particularly for the validation and testing sets, we engage three annotators to verify each video.

Video Captioning The video captioning task requires the model to generate a concise sentence describing a video clip's content and title. To create the dataset, we randomly sample around 80,000 videos based on category frequency distribution and employ a color histogram-based approach for segmenting each video into shots (Mei et al., 2014). To ensure an accurate understanding of the video content and produce precise descriptions, we engage several annotators who are native Chinese speakers with strong educational backgrounds. As part of the pre-annotation process, we assign 25 random videos to each annotator, requesting them to create captions that include the subject and object in the video, as well as relevant descriptions of actions and background. The captions must consist of at least 15 Chinese characters. Following the pre-annotation stage, annotators proceed with annotating the datasets and split them into the training, validation, and testing sets. Especially, to prevent data leakage, clips from the same video or sharing the same title are exclusively assigned to either the training or testing sets. Moreover, for the validation and testing datasets, we enlist more than three individuals to annotate the video clips, promoting diversity and quality.

Video-Text Retrieval Similar to the annotation procedures video captioning task, we first segment the video into clips using a color histogram-based method. Then, these video clips are assigned to different native Chinese speakers for labeling the clips. We also adopt the two-step verification procedure in which each collected description must be reviewed. In addition, we ensure that clips from the same video or those with identical text titles are not exclusively included in the training or test set to prevent potential data leakage.

4 METHODOLOGY

Since the pre-trained large language model shows incredible zero-shot and generalization abilities on various tasks, we use the off-the-shelf Chinese large language model (e.g. GPT-3 (Brown et al., 2020)) for efficient modularized training. To this end, we propose mPLUG-video, a decoder-only based video-language model that leverages the frozen large language model. Specifically, our model consists of a video encoder, a visual abstractor module, and a language decoder, as illustrated in Figure 4. Besides, we only train the video encoder and visual abstractor containing limited parameters , which reduces the computation burden significantly.

4.1 ARCHITECTURE

The Video Encoder We leverage a 12-layer TimeSformer (Bertasius et al., 2021) to extract the video features, with 224×224 input frames. We sparsely sample T frames from each video $\mathcal{V}$, where the TimeSformer first divides the video frames into N non-overlapping patches and flattens them into a sequence of $T \times N$ patches. Then these patches are fed into the patch projection layers for patch representation. To encode the position of each patch, we add learnable embeddings to encode each patch's spatial and temporal position. Then the TimeSformer applies divided spatiotemporal attention to yield video representation $V \in \mathbb{R}^{(T\times N)\times D}$, where D is the hidden dimension of the video representation.

Visual Abstractor Module To mitigate the computation burden with the lengthy video sequences, we introduce visual abstractor module which utilizes learnable queries $Q \in \mathbb{R}^{M\times D}$ for reducing the length of video sequence as follows:

$$\tilde{Q} = CrossAttention(Q, V, V), \tag{1}$$

$$\tilde{Q} = FFN(\tilde{Q}) + \tilde{Q}, \tag{2}$$

where $CrossAttention(x, y, z)$ is the cross-attention layer with Query x, Key y, and Value z. The $FFN(\cdot)$ is the feed-forward layer (Vaswani et al., 2017). Finally, we obtain the reduced video sequence $\tilde{Q} \in \mathbb{R}^{M\times D}$.

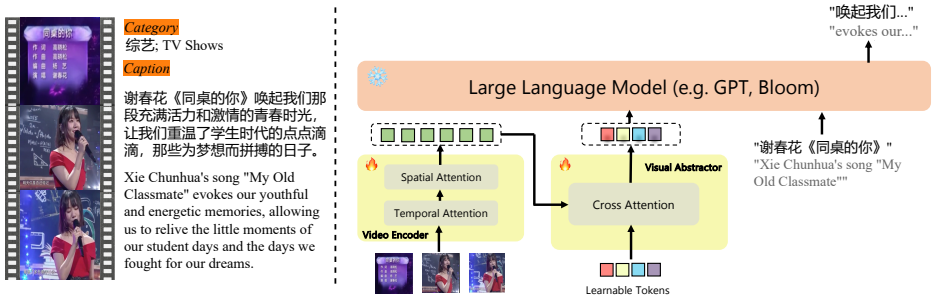


Figure 4: The overview of mPLUG-video.

The Language Decoder Since pre-trained large language models demonstrate strong zero-shot capabilities in text generation, we utilize them as the general text decoder for multi-modal inputs while keeping it frozen. In specific, we treat the video as a foreign language and concatenate the reduced video sequence with the text token features obtained from the text embedding layer. Then, the video and text token features are jointly fed into the large language model which is frozen for obtaining the video-guided language features. Finally, the video-guided language features are predicted for text tokens.

Training Objective We train mPLUG-video within an auto-regressive manner and adopt the next token prediction task for training. In detail, the model needs to complete the texts based on the given video, and the language modeling loss is calculated as:

$$\mathcal{L} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\left[\sum_{l=1}^{L}\log p(w_l|\mathcal{W}_{[0,l)},\mathcal{V})\right], \tag{3}$$

where L denotes the total number of words in the text, and $\mathcal{W}$ denotes the word tokens.

4.2 Application to Downstream Tasks

Video Captioning Video captioning is considered an auto-regressive task. During the process of fine-tuning a video captioning dataset, the training objective remains the same as pre-training.

Video Category Classification We treat video category classification as a video caption task. Annotated category names of videos are regarded as ground-truth captions. We evaluate the accuracy of predictions based on whether the predicted category name exactly matches the ground-truth.

Video-Text Retrieval In contrast to mPLUG-2, which includes a contrastive head and a matching head for the retrieval task, our mPLUG-video cannot be directly used for retrieval tasks. Therefore, we input video-text pairs into the model and extract the feature of the last token. We obtain the matching score by applying an extra linear layer to the feature of the last token.

5 Experiments

5.1 Implementation Details

mPLUG-video leverages the pre-trained Chinese GPT-3 [2] [3] as the language decoder, and the video encoder is pre-trained on ImageNet (Ridnik et al., 2021). During pre-training, we sparsely sample 8 frames from each video preserving their order in-between, and resize them to 224 × 224. We use a batch size of 512 and train mPLUG-video for 10 epochs. We adopt the AdamW optimizer with $\beta = (0.9, 0.98)$, and set the learning rate and weight decay to 1e-4 and 1e-3 respectively. We warm up the training with 2000 warm-up steps then decay the learning rate with the cosine schedule. For downstream tasks, we use a batch size of 128 and train mPLUG-video for 10 epochs with a learning rate of 2e-5.

[2] https://modelscope.cn/models/damo/nlp_gpt3_text-generation_1.3B/summary
[3] https://modelscope.cn/models/damo/nlp_gpt3_text-generation_2.7B/summary

5.2 EVALUATION ON DOWNSTREAM TASKS

In this subsection, we evaluate the performance of ALPRO, mPLUG-2, and mPLUG-video on video category classification, video captioning, and video-text retrieval, respectively.

Evaluation on Video Category Classification We assess the performance of ALPRO, mPLUG-2, and mPLUG-video on video category classification tasks. We measure the top-1 and top-5 accuracy of each model. For the generation models, a generated category name that is exactly the same as ground truth can be regarded as a correct prediction. The comparison results are shown in Table 4. Our results reveal that mPLUG-video achieves the highest accuracy, with a top-1 accuracy of 80.57% and a top-5 accuracy of 98.15%. Interestingly, mPLUG-video (2.7B) outperforms mPLUG-video (1.3B), highlighting the importance of natural language understanding with a larger LLM decoder. Besides, mPLUG-video outperforms the other two models by utilizing the internal knowledge within LLM, showing the effectiveness of decoder-only architecture.

Evaluation on Video Caption We present in Table 4 the performance of models on Video Caption. ALPRO does not have a decoder module. Therefore, its performance was not reported. The performance of mPLUG-Video and mPLUG-2 are compared based on various metrics, including METEOR, ROUGE, CIDEr, and BLEU-4. It is found that mPLUG-video (2.7B) achieves higher scores than mPLUG-Video (1.3B) across all four metrics. Additionally, mPLUG-video obtains higher scores than mPLUG-2 on BLEU-4. These results suggest that pre-trained language models are essential and video captioning tasks based on our dataset are still challenging for existing methods We also present the results on VATEX (Wang et al., 2019) dataset in Table 5, which demonstrates models can benefit from pre-training on Youku-mPLUG.

Evaluation on Video-Text Retrieval Table 6 presents the performance comparison between models on video retrieval task. We observe that mPLUG-2 outperforms ALPRO, possibly due to the incorporation of universal layers that remove modality differences and generate superior uni-modal representations. We also notice that mPLUG-video performs poorly on video retrieval task. Since we only adopt language modeling as the pre-training task, it does not explicitly contain the video-language alignment with contrastive learning.

Table 4: Comparison results on Youku-mPLUG. Video category prediction and video captioning, respectively. For video category prediction, top-1 and top-5 accuracy are reported. For video captioning, we report BELU-4, METEOR, ROUGE, and CIDEr. * denotes the language model is frozen.

	Video Category Prediction		Video Captioning			
Model	Top-1 Acc.(%)	Top-5 Acc.(%)	BLEU-4	METEOR	ROUGE	CIDEr
ALPRO	78.15	95.15	-	-	-	-
mPLUG-2	77.79	92.44	43.7	**27.6**	52.9	67.7
mPLUG-Video (1.3B)*	80.04	98.06	46.4	26.5	52.9	67.7
mPLUG-Video (2.7B)*	**80.57**	**98.15**	**47.1**	26.7	**53.3**	**68.9**

Table 5: Comparison of results on VATEX of Video Captioning.

Model	BLEU-4	METEOR	ROUGE	CIDEr
mPLUG-2	53.6	31.0	59.9	87.0
mPLUG-Video (1.3B w/o pre-train)*	49.2	29.4	58.1	76.8
mPLUG-Video (1.3B w/ pre-train)*	**57.4**	**31.6**	**62.2**	**97.2**

5.3 ABLATION STUDY ON MODALITIES

In this section, we investigate the contributions of different modalities to video-language modeling by leveraging the category classification task on our Youku-mPLUG. Table 7 presents the performance of the baseline model (ALPRO) trained with data of different modalities. Vision Modality and Language Modality denote the model trained with the corresponding modality of data (video frames or video captions). Youku-mPLUG Pre-Trained refers to the model pre-trained on Youku-mPLUG before fine-tuning. The results show that the performance of the model trained with the visual modality

Table 6: Comparison results on Youku-mPLUG. Video retrieval. We evaluate models on video retrieval (V2T) and text retrieval (T2V). we report the average of R@1, R@5 and R@10. * denotes the language model is frozen.

Model	Video Retrieval					
	V2T			T2V		
	R@1	R@2	R@10	R@1	R@5	R@10
ALPRO	27.00	53.33	64.09	26.63	53.20	64.43
mPLUG-2	**38.45**	**65.48**	**75.18**	**38.45**	**65.48**	**75.18**
mPLUG-Video (1.3B)*	7.01	20.33	29.67	7.01	20.33	29.67
mPLUG-Video (2.7B)*	7.62	21.24	31.39	7.62	21.24	31.39

Table 7: Comparison of different modalities and Youku-mPLUG on category classification task.

Vision Modality	Language Modality	Youku-mPLUG Pre-Trained	Top-1 Acc.(%)	Top-5 Acc.(%)
✓	✗	✗	63.51	89.89
✗	✓	✗	59.31	86.31
✓	✓	✗	69.40	90.07
✓	✓	✓	**78.15**	**95.15**

is higher than that with the language modality. This suggests that high-level language modalities may lose fine-grained visual clues, leading to failure in classification. Additionally, we observe that the model trained with both vision and language modalities achieves higher performance than unimodal models. This observation demonstrates the importance of modality complementarity in video understanding. Pre-training the model with Youku-mPLUG leads to a significant improvement in performance, emphasizing the importance of our Youku-mPLUG.

5.4 Human Evaluation of Zero-shot Video Instruction Understanding

To test the video instruction understanding ability of different models, we manually set 65 instructions based on 50 randomly-sampled videos (45 from Youku-mPLUG, 5 from HD-VILA-100M (Xue et al., 2022)). We compare the instruction understanding performance of three models: VideoLLaMA(Zhang et al., 2023), mPLUG-Video w/o pretrain and mPLUG-Video. VideoLLaMA is trained with visual instruction data from MiniGPT-4(Zhu et al., 2023), LLaVa (Liu et al., 2023) and Video-Chat (Li et al., 2023b), while the latter two models only utilize visual training data from LLaVa (Liu et al., 2023). We ask human annotators to score the models' responses. Following Self-Instruct(Wang et al., 2022), human annotators are asked to rate the response into four levels, where A means 'correct and satisfying response', B means 'acceptable response with minor imperfections', C means 'response to the instruction but has significant errors' and D means 'irrelevant or invalid response'. As shown in Fig. 5, with the pertaining on Youku-mPLUG, mPLUG-video achieves much better video instruction understanding and responding ability, demonstrating the effectiveness of our proposed pretraining data. Qualitative results can be found in the supplementary material.

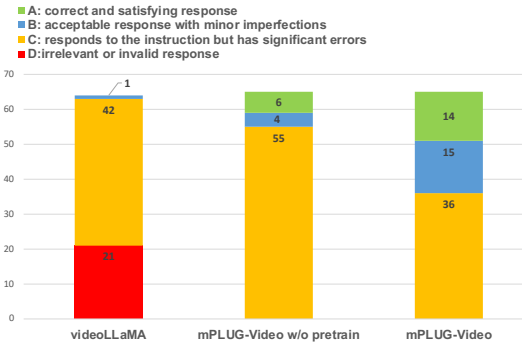


Figure 5: Human evaluation about zero-shot video instruction understanding on 65 cases.

6 Conclusion

In this paper, we introduce the largest high-quality video-language dataset in Chinese, called Youku-mPLUG. Additionally, we present a human-annotated benchmark that comprises three downstream tasks, i.e., Video-Text Retrieval, Video Captioning, and Video Category Classification. We propose a decoder-only model, mPLUG-video, that is modularized and pre-trained on Youku-mPLUG. Results from our experiments indicate that our evaluation set can effectively evaluate the video language comprehension and modeling abilities of models. Furthermore, pre-training on Youku-mPLUG leads to significant improvements, and our mPLUG-video achieves a new state-of-the-art performance.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198, 2022.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In Proceedings of the IEEE international conference on computer vision, pp. 5803–5812, 2017.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1728–1738, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In ICML, pp. 4, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp. 190–200, 2011.

Tian Gan, Qing Wang, Xingning Dong, Xiangyuan Ren, Liqiang Nie, and Qingpei Guo. Cnvid-3.5m: Build, filter, and pre-train the large-scale public chinese video-text dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14815–14824, June 2023.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16167–16176, 2022.

Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, December 2020.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2758–2766, 2017.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In International Conference on Computer Vision (ICCV), 2017.

Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (eds.), MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, pp. 2567–2576. ACM, 2021a. doi: 10.1145/3474085.3475431. URL https://doi.org/10.1145/3474085.3475431.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7331–7341, 2021b.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. arXiv preprint arXiv:2205.12005, 2022a.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4953–4963, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2023b.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. In Conference on Empirical Methods in Natural Language Processing, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. CoRR, abs/2304.08485, 2023.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. ArXiv, abs/2002.06353, 2020.

Avinash Madasu, Estelle Aflalo, Gabriela Ben-Melech Stan, Shao-Yen Tseng, Gedas Bertasius, and Vasudev Lal. Improving video retrieval using multilingual knowledge transfer. In European Conference on Information Retrieval, 2022.

Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. Multimedia search reranking: A literature survey. ACM Comput. Surv., 46:38:1–38:38, 2014.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In ICCV, pp. 2630–2640, 2019.

Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and A. Bimbo. Search-oriented micro-video captioning. Proceedings of the 30th ACM International Conference on Multimedia, 2022a.

Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. Search-oriented micro-video captioning. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, pp. 3234–3243, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548180. URL https://doi.org/10.1145/3503161.3548180.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972, 2021.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3202–3212, 2015.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 4580–4590. IEEE, 2019. doi: 10.1109/ICCV.2019.00468. URL https://doi.org/10.1109/ICCV.2019.00468.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560, 2022.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko,

Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM international conference on Multimedia, pp. 1645–1653, 2017.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video, 2023.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5288–5296, 2016.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In International Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. ArXiv, abs/2211.01335, 2022.

Xu Yang, Zhangzikang Li, Haiyang Xu, Hanwang Zhang, Qinghao Ye, Chenliang Li, Ming Yan, Yu Zhang, Fei Huang, and Songfang Huang. Learning trajectory-word alignments for video-language tasks, 2023.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training, 2022.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. Advances in Neural Information Processing Systems, 34:23634–23651, 2021.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. Poet: Product-oriented video captioner for e-commerce. In Proceedings of the 28th ACM International Conference on Multimedia, MM '20, pp. 1292–1301, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413880. URL https://doi.org/10.1145/3394171.3413880.

Ziqi Zhang, Yuxin Chen, Zongyang Ma, Zhongang Qi, Chunfeng Yuan, Bing Li, Ying Shan, and Weiming Hu. Create: A benchmark for chinese short video retrieval and title generation, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

A Qualitative Results of Zero-shot Video Instruction Understanding

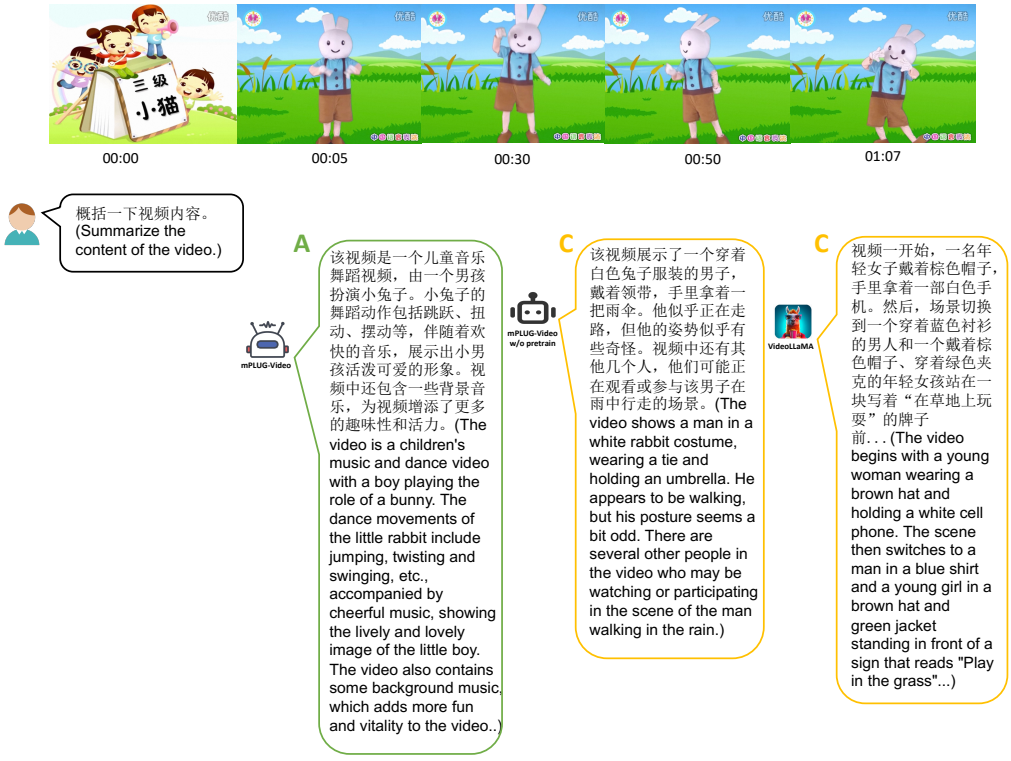


Figure 6: Qualitative results about video captioning.

In this section, we show some qualitative results of zero-shot video instruction understanding to validate the effectiveness of pretraining with Youku-mPLUG. Firstly, as shown in Fig. 6, with pertaining, mPLUG-video can better understand both overall video semantic ('dance video') and detailed vision information, such as 'jumping' and 'twisting'. Besides, Fig. 7 shows that mPLUG-video can better recognize scene texts in videos and give more accurate responses. Therefore, pretraining with Youku-mPLUG is highly likely to enhance the models' ability to comprehend scene text in videos. Furthermore, in Fig. 8, mPLUG-video could accurately point out the key character name 'Ultraman', while the other two models can't. This validates the contribution of pretraining for learning open-domain knowledge.

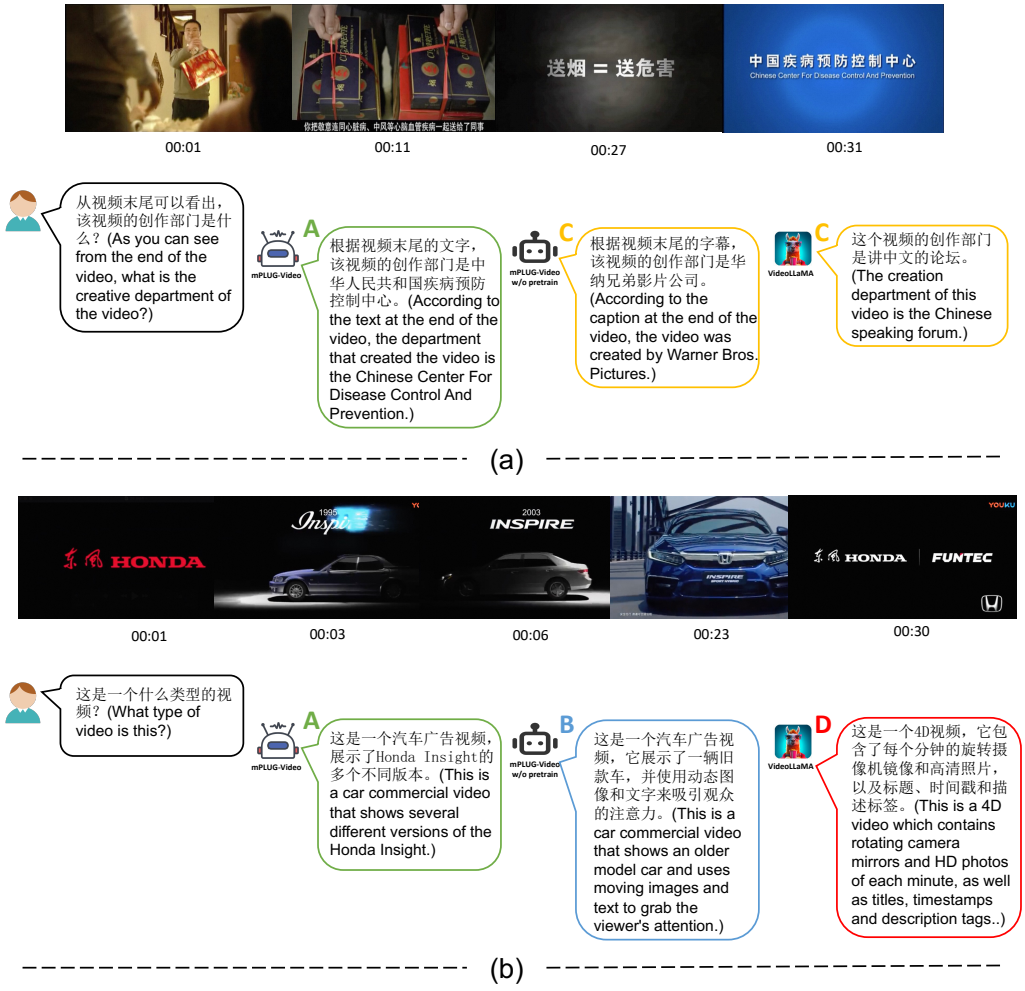


Figure 7: Qualitative results about video scene text understanding.

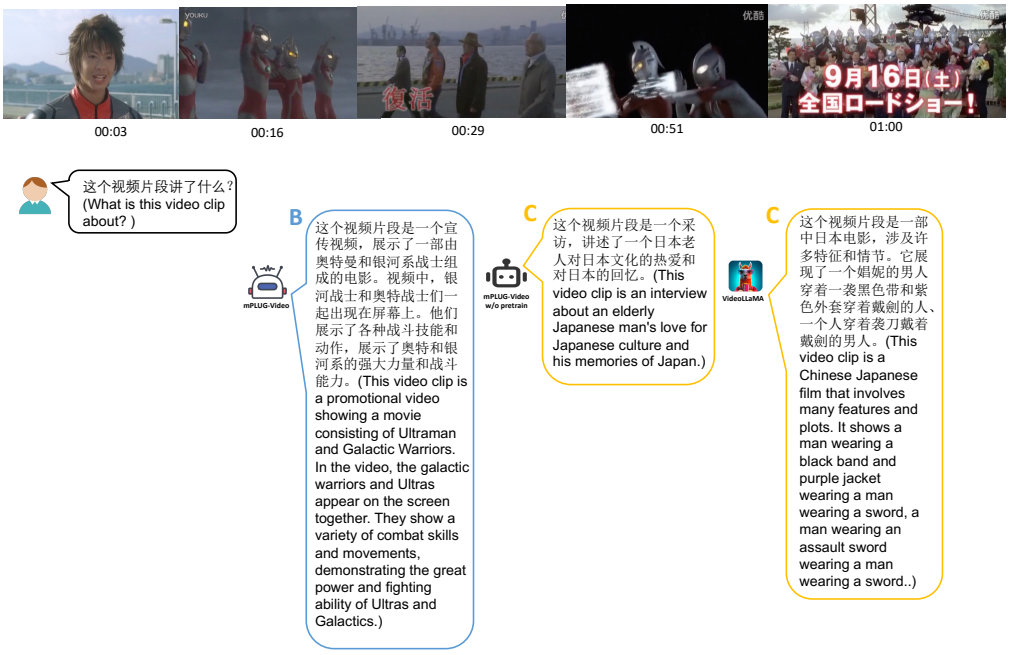


Figure 8: Qualitative results about open-domain knowledge understanding.

B Limitations and Societal Impacts

The Youku-mPLUG dataset predominantly contains concepts and language expressions that were current at the time of collection. As language evolves alongside human activities, our dataset may not encompass emerging concepts, words, and language expressions in the future. This limitation applies to image data as well, where new visual objects or designs might not be captured. Nevertheless, this issue can be addressed by fine-tuning pre-trained models on up-to-date data. Additionally, our dataset is constructed using corpora from the Chinese Internet, meaning the vocabulary and expressions may largely align with Chinese culture. Furthermore, our dataset lacks very long texts and lengthy videos, potentially limiting the ability of the pre-trained models to understand extensive content such as full-length movies.

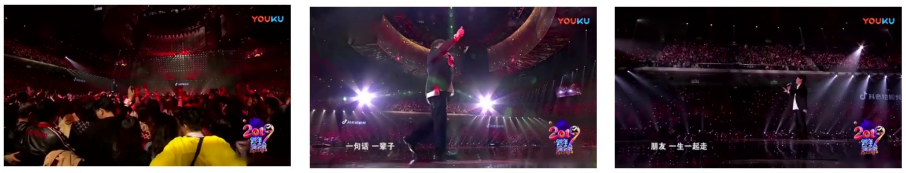
C Hosting, Maintenance Plan, and License

Long-term maintenance of Youku-mPLUG and models proposed and evaluated in our paper will be made by the authors. The dataset will be hosted on the Modelscope[4] with Alibaba Cloud as the backend for the downloading service. For stability, we would check the URLs of the dataset regularly and fix those broken videos in time. Our released datasets are provided under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("CC BY-NC-SA 4.0"), with the additional terms included herein[5]. When users download or use the datasets from our website, they must agree to the license.

[4]https://www.modelscope.cn/

[5]https://creativecommons. org/licenses/by-nc-sa/4.0/legalcode

阿米尔汗直播打麻将 王宝强调侃米叔是 “你输” —早班机
Amir Khan plays mahjong live and Wang Baoqiang jokes that "you lose" to Amir Khan - Morning Flight



治愈系旋律来袭，周华健经典再现《朋友》，珍惜身边的人吧！
Healing melody strikes, Zhou Huajian classic reproduction of "Friends", cherish the people around you!

刘宇宁你把手撒开让我来
Liuyuning, let go of your hand and let me do it.

【江河水 第22集】秦昊危机关头舍己为人勇救阚清子尽显真情！
At the crisis point in River of Rivers Episode 22, Qin Hao showed true affection by riski
ng his life to save Kan Qingzi!

又帅气又简单的橡皮筋魔术1 (校园魔术)
Handsome and Simple Rubber Band Magic 1 (Campus Magic)

Figure 9: Examples in Youku-mPLUG.